Centre for
Data Analytics

Insight

Matrix Factorization For Topic Models

Dr. Derek Greene
Insight Latent Space Workshop

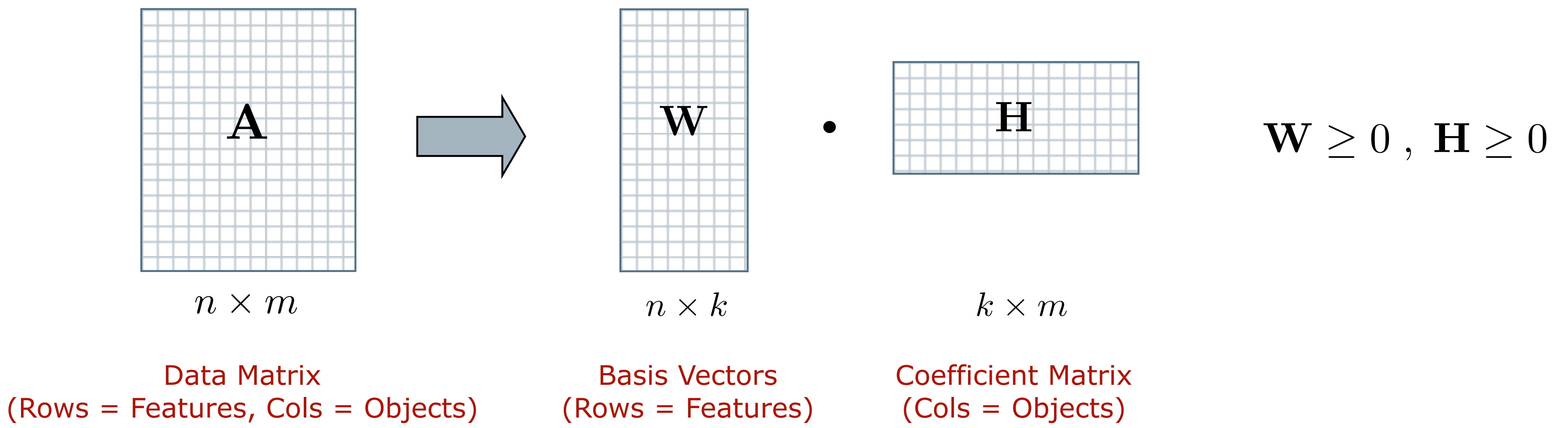
DCU NUI Galway OÉ Gaillimh UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh UCD DUBLIN sfi Science Foundation Ireland

Non-negative Matrix Factorization

- *NMF*: an unsupervised family of algorithms that simultaneously perform dimension reduction and clustering.
- Also known as *positive matrix factorization* (PMF) and *non-negative matrix approximation* (NNMA).
- No strong statistical justification or grounding.
- But has been successfully applied in a range of areas:
  - Bioinformatics (e.g. clustering gene expression networks).
  - Image processing (e.g. face detection).
  - Audio processing (e.g. source separation).
  - Text analysis (e.g. document clustering).

NMF Overview

- NMF produces a "parts-based" decomposition of the latent relationships in a data matrix.
- Given a non-negative matrix **A**, find *k*-dimension approximation in terms of non-negative factors **W** and **H** (Lee & Seung, 1999).



- Approximate each object (i.e. column of **A**) by a linear combination of *k* reduced dimensions or "basis vectors" in **W**.
- Each basis vector can be interpreted as a cluster. The memberships of objects in these clusters encoded by **H**.

NMF Algorithm Components

- **Input:** Non-negative data matrix (**A**), number of basis vectors (*k*), initial values for factors **W** and **H** (e.g. random matrices).
- **Objective Function:** Some measure of reconstruction error between **A** and the approximation **WH**.

Euclidean Distance (Lee & Seung, 1999)

$$\frac{1}{2}\,||\mathbf{A}-\mathbf{WH}||_{\mathsf{F}}^{2} = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(A_{ij}-(WH)_{ij}\right)^{2}$$

- **Optimisation Process:** Local EM-style optimisation to refine **W** and **H** in order to minimise the objective function.
- Common approach is to iterate between two multiplicative update rules until convergence (Lee & Seung, 1999).

1. Update **H**

$$H_{cj} \leftarrow H_{cj}\,\frac{(W^{\mathsf{T}}\mathbf{A})_{cj}}{(W^{\mathsf{T}}\mathbf{WH})_{cj}}$$

2. Update **W**

$$W_{ic} \leftarrow W_{ic}\,\frac{(\mathbf{A}H^{\mathsf{T}})_{ic}}{(\mathbf{WH}H^{\mathsf{T}})_{ic}}$$

NMF Variants

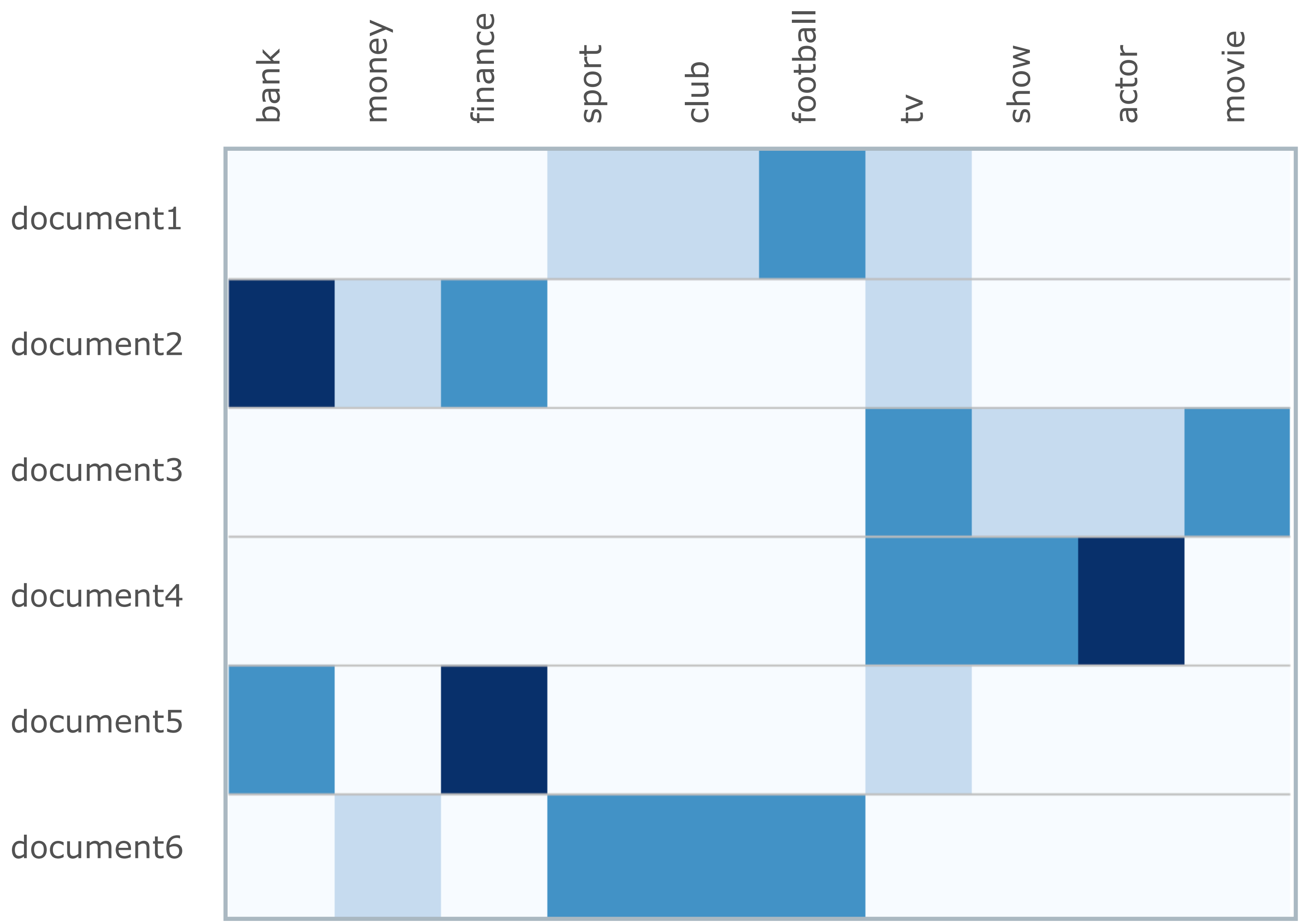
- **Different objective functions:**
  - KL divergence; Bregman divergences (Sra & Dhillon, 2005).
- **More efficient optimisation:**
  - Alternating least squares with projected gradient method for sub-problems (Lin, 2007).
- **Constraints:**
  - Enforcing sparseness in outputs (e.g. Liu et al, 2003).
  - Incorporation of background information (Semi-NMF).
- **Different inputs:**
  - Symmetric matrices - e.g. document-document cosine similarity matrix (Ding & He, 2005).

Application: Topic Models

- **Recommended methodology:**
  1. Construct vector space model for documents (after stop-word filtering), resulting in a term-document matrix **A**.
  2. Apply TF-IDF term weight normalisation to **A**.
  3. Normalize TF-IDF vectors to unit length.
  4. Initialise factors using NNDSVD on **A**.
  5. Apply Projected Gradient NMF to **A**.
- **Interpreting NMF output:**
  - *Basis vectors:* the topics (clusters) in the data.
  - *Coefficient matrix*: the membership weights for documents relative to each topic (cluster).

NMF Topic Modeling: Simple Example

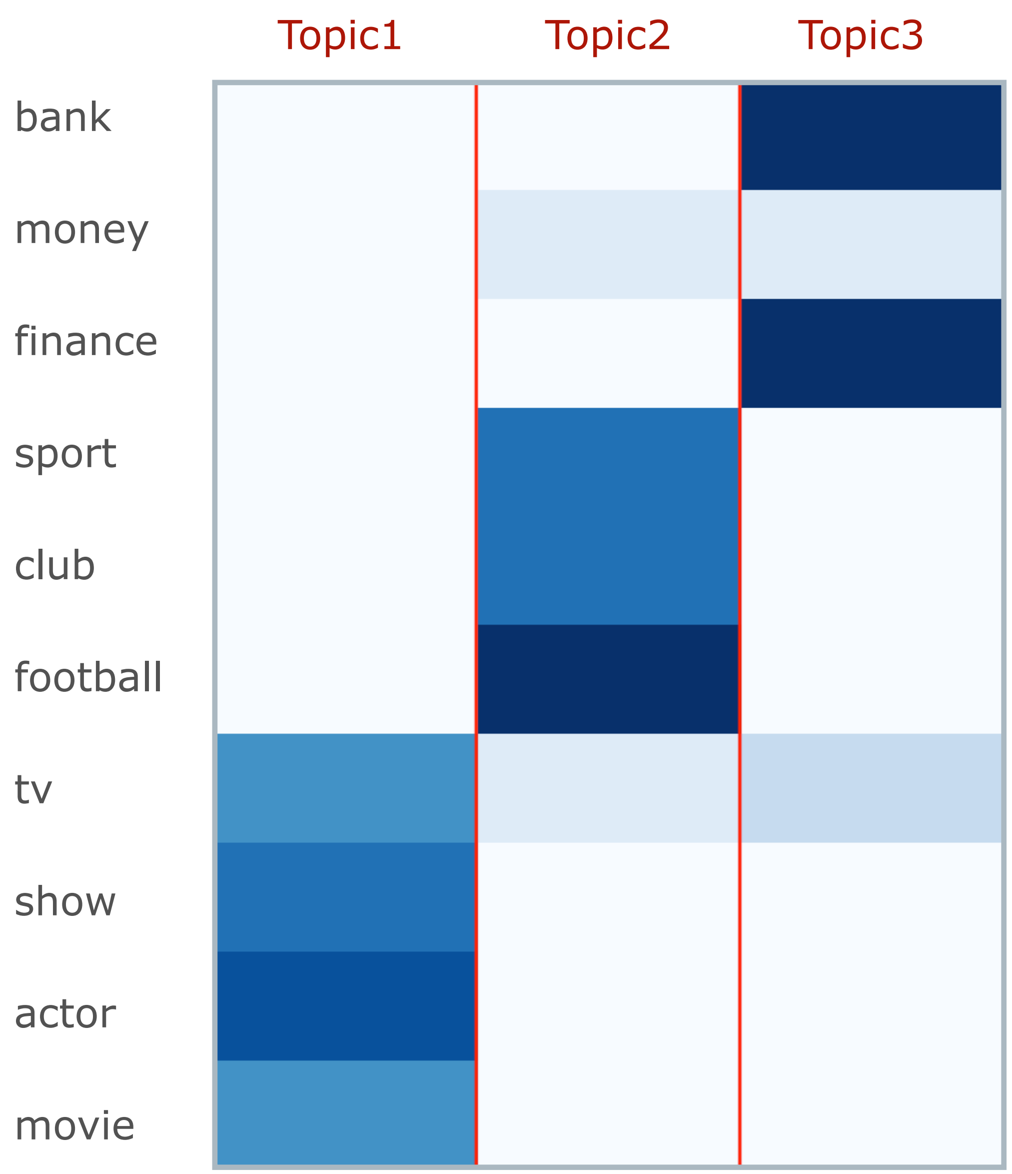
Document-Term Matrix **A**
(6 rows x 10 columns)



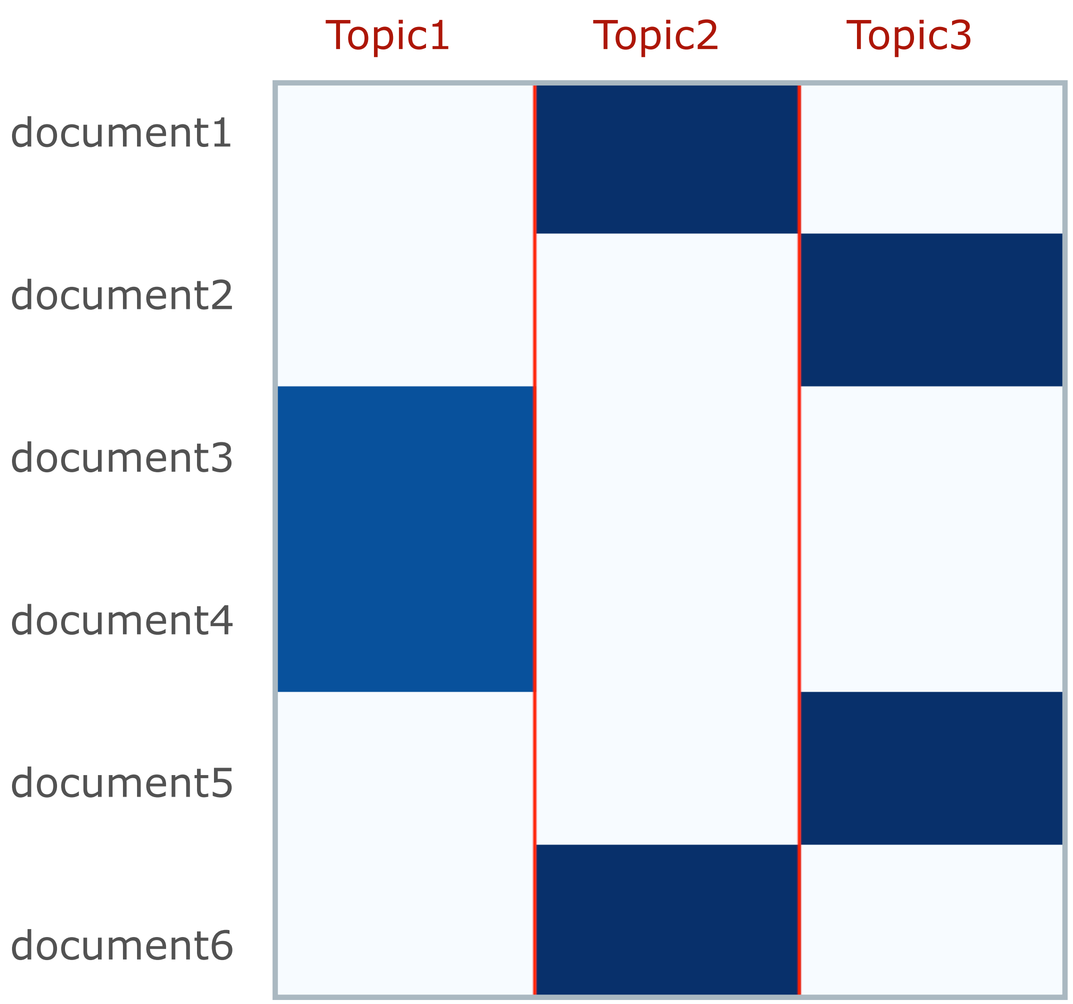
- Apply TF-IDF and unit length normalization to rows of **A**.
- Run Euclidean NMF on normalized **A** (k=3, random initialization).

NMF Topic Modeling: Simple Example

Basis vectors **W**: topics (clusters)



Coefficients **H**: memberships for documents



Challenge: Selecting *K*

- As with LDA, the selection of number of topics *k* is often performed manually. No definitive model selection strategy.
- Various alternatives comparing different models:
  - Compare reconstruction errors for different parameters. Natural bias towards larger value of *k*.
  - Build a “consensus matrix” from multiple runs for each *k*, assess presence of block structure (Brunet et al, 2004).
  - Examine the *stability* (i.e. agreement between results) from multiple randomly initialized runs for each value of *k*.

Challenge: Algorithm Initialization

- Standard random initialisation of NMF factors can lead to *instability* - i.e. significantly different results for different runs on the same data matrix.
- *NNDSVD*: Nonnegative Double Singular Value Decomposition (Boutsidis & Gallopoulos, 2008):
    - Provides a deterministic initialization with no random element.
    - Chooses initial factors based on positive components of the first k dimensions of SVD of data matrix **A**.
    - Often leads to significant decrease in number of NMF iterations required before convergence.

Experiment: BBC News Articles

- Collection of 2,225 BBC news articles from 2004-2005 with 5 manually annotated topics (http://mlg.ucd.ie/datasets/bbc.html).
- Applied Euclidean Projected Gradient NMF (k=5) to 2,225 x 9,125 matrix.
- Extract topic "descriptions" based on top ranked terms in basis vectors.

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
growth	mobile	england	film	labour
economy	phone	game	best	election
year	music	win	awards	blair
bank	technology	wales	award	brown
sales	people	cup	actor	party
economic	digital	ireland	oscar	government
oil	users	team	festival	howard
market	broadband	play	films	minister
prices	net	match	actress	tax
china	software	rugby	won	chancellor

Experiment: Irish Economy Dataset

- Collection of 21k news articles from 2009-2010 relating to the economy (Irish Times, Irish Independent & Examiner).
- Extracted all named entities from articles (person, org, location), and constructed 21,496 x 3,014 article-entity matrix.
- Applied Euclidean Projected Gradient NMF (k=8) matrix.

Topic 1	Topic 2	Topic 3	Topic 4
nama	european_union	allied_irish_bank	hse
brian_lenihan	europe	bank_of_ireland	dublin
green_party	greece	anglo_irish_bank	mary_harney
ntma	lisbon_treaty	dublin	department_of_health
anglo_irish_bank	ecb	irish_life_permanent	brendan_drumm

Topic 5	Topic 6	Topic 7	Topic 8
usa	aer_lingus	uk	brian_cowen
asia	ryanair	dublin	fine_gael
new_york	dublin	northern_ireland	fianna_fail
federal_reserve	daa	bank_of_england	green_party
china	christoph_mueller	london	brian_lenihan

Experiment: IMDb Dataset

- Constructed documents from IMDb Keywords for set of 21k movies (http://www.imdb.com/Sections/Keywords/).
- Applied NMF (k=10) to 20,923 x 5,528 movie-keyword matrix.
- Topic “descriptions” based on top ranked keywords in basis vectors appear to reveal genres and genre cross-overs.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
cowboy	bmovie	martialarts	police	superhero
shootout	atgunpoint	combat	detective	basedoncomic
cowboyhat	bwestern	hero	murder	superheroine
cowboyboots	stockfootage	actionhero	investigation	dccomics
horse	gangmember	brawl	policedetective	secretidentity
revolver	duplicity	fistfight	detectiveseries	amazon
sixshotter	gangleader	disarming	murderer	culttv
outlaw	deception	warrior	policeofficer	actionheroine
rifle	sheriff	kungfu	policeman	twowordtitle
winchester	povertyrow	onemanarmy	crime	bracelet

Experiment: IMDb Dataset

- Constructed documents from IMDb Keywords for set of 21k movies (http://www.imdb.com/Sections/Keywords/).
- Applied NMF (k=10) to 20,923 x 5,528 movie-keyword matrix.
- Topic "descriptions" based on top ranked keywords in basis vectors appear to reveal genres and genre cross-overs.

Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
worldwartwo	monster	love	newyorkcity	shotinthechest
soldier	alien	friend	manhattan	shottodeath
battle	cultfilm	kiss	nightclub	shotinthehead
army	supernatural	adultery	marriageproposal	punchedintheface
1940s	scientist	infidelity	jealousy	corpse
nazi	surpriseending	restaurant	engagement	shotintheback
military	demon	extramaritalaffair	party	shotgun
combat	occult	photograph	hotel	shotintheforehead
warviolence	possession	tears	deception	shotintheleg
explosion	slasher	pregnancy	romanticrivalry	shootout

Implementations of NMF

- *Scikit-learn* ML library for Python (http://scikit-learn.org/)
- Implementation of vanilla NMF with Euclidean objective and Projected Gradient for sparse & dense data.

```
from sklearn import decomposition
model = decomposition.NMF(n_components=5, max_iter=100)
result = model.fit(X)
print result.components_
```

- More comprehensive and efficient implementations for NMF variants in Python *NIMFA* package (http://nimfa.biolab.si/)
- R package (http://cran.r-project.org/web/packages/NMF/)
- Also C & MATLAB implementations optimised to use FORTRAN linear algebra libraries & GPUs.

References

- D.D. Lee & H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–91, 1999.
- C. Lin. Projected gradient methods for non-negative matrix factorization. *Neural Computation*, 19(10):2756–2779, 2007
- S. Sra & I.S. Dhillon. Generalized nonnegative matrix approximations with bregman divergences. In *Proc. Advances in Neural Information Processing Systems (NIPS'05)*, 2005.
- Liu, W., Zheng, N. & Lu, X. Non-negative matrix factorization for visual coding. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, vol. 3, 2003.
- C. Ding & X. He. On the Equivalence of Non-negative Matrix Factorization and Spectral Clustering. In *Proc. SIAM International Conference on Data Mining (SDM'05)*, 2005.
- J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proc. National Academy of Sciences,* 101(12):4164–4169, 2004.
- C. Boutsidis & E. Gallopoulos. SVD based initialization: A head start for non-negative matrix factorization. *Pattern Recognition*, 2008.